The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21)

Skills2Job: A Recommender System that Encodes Job Offer Embeddings on Graph Databases (Student Abstract)

Andrea Seveso [2], Anna Giabelli[2], Lorenzo Malandri[1], Fabio Mercorio [1], Mario Mezzanzanica [1]
[1] Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Milan
[2] Department of Informatics, Systems and Communication, University of Milano-Bicocca, Milan
{andrea.seveso, anna.giabelli, lorenzo.malandri, fabio.mercorio, mario.mezzanzanica}@unimib.it

Abstract

We propose a recommender system that, starting from a set of users skills, identifies the most suitable jobs as they emerge from a large text of Online Job Vacancies (OJVs). To this aim, we process 2.5M+ OJVs posted in three different countries (United Kingdom, France and Germany), generating several embeddings and performing an intrinsic evaluation of their quality. Besides, we compute a measure of skill importance for each occupation in each country, the Revealed Comparative Advantage (*rca*). The best vector models, together with the *rca*, are used to feed a graph database, which will serve as the keystone for the recommender system. Finally, a user study of 10 validates the effectiveness of `skills2job`, both in terms of precision and nDGC.

Introduction

Given the very high number of job positions and applicants on online job portals, the problem of person-job fit has become relevant in recent literature, both as a skill measuring system (Xu et al. 2018) and job recommendation system (Zhang et al. 2016). Recommender systems in the labour market domain rely strongly on handcrafted features and expert knowledge, which make them costly, difficult to update and error-prone. For that reason, we propose `skills2job`, a knowledge poor and data driven job recommendation system, which can be adapted to different countries/industries and easily updated over time. Moreover, `skills2job` is the first system that organises labour market information, automatically extracted from a large corpus of OJVs, in a graph database which can be queried to recommend the most suitable occupations for an user based on its skills. `skills2job` uses ESCO (https://ec.europa.eu/esco) as a target taxonomy to organise occupations and skills, to allow querying the graph using 27 languages. `skills2job` was realised as part of the research activity of an EU project[1] (see (Boselli et al. 2017)), which aims at realising the first EU real-time labour market monitor, by collecting and classifying OJVs from all 27+1 EU countries.

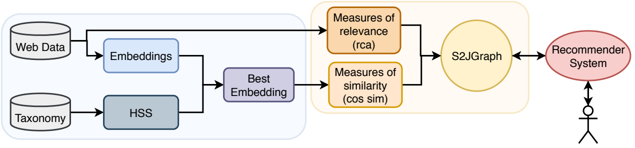

Figure 1: Workflow of steps for building `skills2job`

An Overview of `skills2job`

The workflow of `skills2job`, presented in Fig. 1, can be divided in five main steps: **S.1** To extract linguistic patterns from OJVs, we train multiple embedding models through FastText, a library for representation learning which builds word embeddings considering sub-word information by representing each word as the sum of its character *n*-gram vectors; **S.2** we compute measure of pairwise semantic similarity between taxonomic elements, namely `HSS` (developed in (Giabelli et al. 2020b) and previously called Hierarchical Semantic Relatedness (*HSR*)).

Compared with `HSS` , previous measures of semantic similarity in taxonomies (see Aouicha, Taieb, and Hamadou (2016) for a survey) suffer of two main limitations. First, when a word has multiple senses, those methods compute a value of similarity for each word sense and then consider only the highest, which is the self-information of the least frequent lowest common ancestor. As a consequence, more specific senses will have a higher value of similarity, but this does not reflect the use of words in advertising job positions; second, though they consider the structure of the taxonomy (i.e., the relationship between concepts) they do not take into account the number of child entities (i.e., words) belonging to those concepts. This is crucial in our case as ESCO includes generic concepts that, in turn, contain many different occupations. On the contrary, some very specific concepts can be represented by a few occupations which are highly informative. The aim of `HSS` is to overcome these limitations to work with the ESCO taxonomy. **S.3** To select the embedding that better preserves taxonomic relations, we perform an intrinsic evaluation by computing the Pearson correlation of the cosine similarity between each couple of skills and their corresponding `HSS`. **S.4** We employ co-occurrence statistics, using a normalized count based measure of skill-relevance, the Revealed Comparative Advan-

[1] CEDEFOP 2014. Real-time Labour Market information on skill requirements: feasibility study and working prototype". https://goo.gl/qNjmrn

Skill s	c_S	c_A
implement front-end website design	0.59	0.60 $\leftrightarrow$
CSS	0.51	0.63 $\uparrow$
C#	0.46	0.40 $\downarrow$
use markup languages	0.17	0.61 $\uparrow\uparrow$
skill gap o_i	rca_{NORM}	
perform online data analysis	1	
social media management	0.75	
social media marketing techniques	0.66	

Table 1: Highest ranked result, "Web Technicians", showing the skills rca_{NORM} in c_S and c_A and the skill gap o_i.

	P@3-3	P@3-4	nDCG
Result	0.823	0.610	0.985

Table 2: User eval results for the two methods. P@3-N indicates a user score of at least N is considered a true positive.

tage (*rca*) (Alabdulkareem et al. 2018). **S.5** The information extracted through word embeddings and *rca* is stored in our graph database, called S2JGraph, which is formalized as a directed labelled multi-graph and the formalization is inspired by (Giabelli et al. 2020a). Note that both the *rca* and the best embedding are computed for each country, capturing the difference between the requested skills and the occupation terms as they are used in different countires.

Skill Based Recommendations

The graph database, S2JGraph, is used as a keystone for several recommendations tasks using the Cypher query language. Given a set of starting skills **S**, a starting occupation o_S, a starting country c_S, a target country c_A and a target skill s_T provided by the user, skills2job returns:

(i) The relevance of each $s \in \mathbf{S}$ for o_S in c_S;

(ii) A list of occupations **O** in c_A and for each $o_i \in \mathbf{O}$: (a) The relevance of each $s \in \mathbf{S}$ for o_i; (b) A list of skills that o_i requires different from those in **S** and relevant for o_i (namely, the *skill gap*).

(iii) A set of skills recommended to the user given **S** and s_T.

The main use case - query (ii) - recommends a series of occupations in a target labor market based on the user's skills, matching all the occupations in the target country c_A which require at least one of the starting skills in **S**. Then the query matches all the skills which are required by the target occupation with a $rca > \alpha$ and which have a *cosine similarity* with all of the starting skills in $\mathbf{S} < \beta$. These are the *skill gap*, which are relevant for the target occupation (high *rca*) and different enough from the starting skills (low *cosine similarity*). Those are skills that she/he should acquire to do that job in the target country.

An example of query (ii) is reported in Tab. 1 ($\alpha = 0.6$, $\beta = 0.7$). The starting parameters are the following: $\mathbf{S}$ =["implement front-end website design", "CSS","C#", "use markup languages"]; o_S ="Web and multimedia developers"; $c_S = UK$: $c_A = DE$; s_T = "Python".

User Evaluation and Conclusion Remark

The results of skills2job were evaluated through a user study following (Kanakia et al. 2019). We asked 10 Labor Market experts belonging to the European Network on Regional Labour Market Monitoring to judge whether the starting skills are relevant for the occupations provided by the system or not, using a Likert scale. As 3 recommendations were presented for each item, we decided to use Precision@3 (P@3), assuming either a user score of at least 3 as a true positive (P@3-3), or of at least 4 (P@3-4). The normalized Discounted Cumulative Gain (nDCG) has also been computed, which measures the usefulness of an item based on its position in the result list. The results (see Tab. 2) show a high degree of correlation between the user evaluation and the recommendation ranking.

In conclusion, skills2job identifies the most suited job on the basis of a set of user's skills, encoding the skill relevance as emerges from real-labour market demand. It can process any OJV dataset in any EU language. Here we used 2.5M+ vacancies processed trough distributional semantics and co-occurrence statistics, organised in a graph database. A user evaluation made by experts show the system is effective in identifying the correct job given a set of user skills.

We have been working to extend skills2job to all 26+1 EU Countries, enabling policy-makers to observe the labour market demand at skill level.

References

Alabdulkareem, A.; Frank, M. R.; Sun, L.; AlShebli, B.; Hidalgo, C.; and Rahwan, I. 2018. Unpacking the polarization of workplace skills. *Science Advances* 4(7).

Aouicha, M. B.; Taieb, M. A. H.; and Hamadou, A. B. 2016. Taxonomy-based information content and wordnet-wiktionary-wikipedia glosses for semantic relatedness. *Applied Intelligence* 45(2): 475–511.

Boselli, R.; Cesarini, M.; Mercorio, F.; and Mezzanzanica, M. 2017. Using Machine Learning for Labour Market Intelligence. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 330–342. Springer.

Giabelli, A.; Malandri, L.; Mercorio, F.; and Mezzanzanica, M. 2020a. GraphLMI: A data driven system for exploring labor market information through graph databases. *Multimedia Tools and Applications* 1–30.

Giabelli, A.; Malandri, L.; Mercorio, F.; Mezzanzanica, M.; and Seveso, A. 2020b. NEO: A Tool for Taxonomy Enrichment with New Emerging Occupations. In *International Semantic Web Conference*, 568–584. Springer.

Kanakia, A.; Shen, Z.; Eide, D.; and Wang, K. 2019. A scalable hybrid research paper recommender system for microsoft academic. In *WWW*, 2893–2899.

Xu, T.; Zhu, H.; Zhu, C.; Li, P.; and Xiong, H. 2018. Measuring the popularity of job skills in recruitment market: A multi-criteria approach. In *AAAI*.

Zhang, X.; Zhou, Y.; Ma, Y.; Chen, B.-C.; Zhang, L.; and Agarwal, D. 2016. Glmix: Generalized linear mixed models for large-scale response prediction. In *SIGKDD*, 363–372.